PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 5/04 — A1

(11) International Publication Number: **WO 99/66496**

(43) International Publication Date: 23 December 1999 (23.12.99)

(21) International Application Number: PCT/US99/13329

(22) International Filing Date: 14 June 1999 (14.06.99)

(30) Priority Data:
09/098,669 17 June 1998 (17.06.98) US

(71) Applicant: ONLINE ANYWHERE [US/US]; Building A, Suite 202, 3145 Porter Drive, Palo Alto, CA 94304 (US).

(72) Inventors: SOCHER, Gudrun; 266 Pamela Drive #12, Mountain View, CA 94040 (US). VISHWANATH, Mohan; 537 Tarter Court, San Jose, CA 95136 (US). MENDHEKAR, Anurag; 946 Tamarack Lane #11, Sunnyvale, CA 94086 (US).

(74) Agents: FLIESLER, Martin, C. et al.; Fliesler, Dubb, Meyer and Lovejoy LLP, Suite 400, Four Embarcadero Center, San Fransisco, CA 94111-4156 (US).

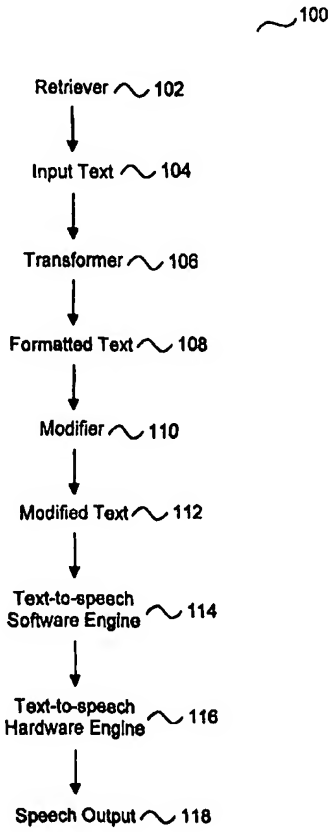
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: INTELLIGENT TEXT-TO-SPEECH SYNTHESIS

(57) Abstract

A method and an apparatus of synthesizing speech from a piece of input text (104). In one embodiment, the method includes the steps of retrieving the input text (104) entered into a computing system, and transforming the input text (104) based on the semantics (152) of at least one word of the input text (104) to generate a formatted text (108) for speech synthesis. In another embodiment, the transformation also depends on at least one characteristic of the person listening to the speech output (118). In yet another embodiment, the transformation further depends on at least one characteristic of the hardware employed by the user to listen to the speech output (118). The transformed text can be further modified to fit a text-to-speech engine to generate the speech output (118).



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTELLIGENT TEXT-TO-SPEECH SYNTHESIS

BACKGROUND OF THE INVENTION

The present invention relates generally to text-to-speech synthesis and more particularly to intelligent text-to-speech synthesis.

We receive a lot of information through hearing, especially when our visual attention is needed for other tasks, such as driving. Radio is a good source of audible documents, and some of us become quite dependent on it. Based on one study, on average, every family in the United States has five radios. Though radio might have become indispensable, the programs put forward by radio stations might not necessarily be what we are currently interested in.

Read-out documents or audio-documents, for example, novels, are available on the market. However, such tapes seem to be only available for a specific market sector. For example, there does not seem to be audio-documents for information with a short lifetime, such as news, weather forecasts or results of sport events. Some information, e.g. stock quotes, is only valuable for a very short period of time, and it would make no sense to produce such audio-documents.

A large number of audio-documents can be produced by automatically translating text into speech output. General discussions of such text-to-speech synthesis systems can be found, for example, in the following publications:

1. *Multilingual Text-to-Speech Synthesis, The Bell Labs Approach*, written by Richard Sproat, and published by Kluwer Academic Publishers, in 1998.

2. IBM ViaVoice.

Such systems typically perform direct word to sound transformation. The speech output is usually not very natural, and they tend to make mistakes. This might be because such systems are not "aware" of what they are reading.

The way we read takes into account what we are reading. For example, if we are reading the topic sentence of a news report, typically, we put in some emphasis. But, since existing systems do not seem to have any clue as to the meaning of the text they are transforming, they tend to transform input texts in the same speed, tone and volume. That is one of the reasons why the speech outputs of existing systems are typically monotonic and boring.

The way we read also should take into account our listener. If our listener is visually impaired and we are describing an object, we should include more details in the object. Moreover, the way we speak should also consider the hardware a listener employs to hear. For example, if your message is heard in a noisy room, probably, you should speak louder.

It should be apparent from the foregoing that there is still a need for an intelligent text-to-speech synthesizer that is, for example, sensitive to the content of the text, sensitive to the one hearing the text or adapts to the hardware the listener employs to hear the text.

SUMMARY OF THE INVENTION

The present invention provides methods and apparatus to synthesize speech from text intelligently. Different important, but previously ignored, factors in the present invention improve on the speech generated. The invented speech synthesizer can take into account the semantics of the input text. For example, if it is a man who should be speaking, a male voice will be used. The synthesizer can take into account the user profile of the person hearing the input text. The synthesizer can also be sensitive to the hardware the user employs to listen to the input text. Thus, the text-to-speech synthesizer is much more intelligent than those in the market.

There are a number of ways to implement the invention. In one embodiment, the synthesizer includes a transformer, a modifier, a text-to-speech software engine and a speech hardware. The transformer analyzes the input text and transforms it into a formatted text. The modifier then modifies this formatted text to fit the requirements of the text-to-speech software engine, whose outputs are fed to the speech hardware to generate the output speech.

The input text has a number of characteristics. It belongs to a class that has at least one specific pattern. For example, the pattern may be that the most important paragraphs of some type of articles are the first one and the last one, as in a newspaper.

The formatted text also has a number of characteristics. It can be independent of the text-to-speech software engine; for example, it is written in Extensible Markup Language (XML).

In one embodiment, the generation of the formatted text is based on the semantics of at least one word of the text. The semantics can be determined by an author--a human being. In another approach, the semantics is generated through mapping the words to a database. For example, if the word is the name of a company, then the database can bring in additional information about the company, such as its stock price at a specific time. In another approach, the semantics is generated through an inference machine. For example, if the words are "Mr. Clinton," the inference machine, based on some pre-stored rules, will assume that the words refer to a male person. Then, a male voice might be used for that purpose.

In another embodiment, the transformation to generate the formatted text is based on at least one characteristic of the user listening to the synthesized speech. In yet another embodiment, the transformation to generate the formatted text depends on at least one characteristic of the hardware the user employs to listen to the synthesized speech. The above embodiments can be mixed and matched. For example, the transformation can be based on semantics of at least one word of the text and one characteristic of the user listening to the synthesized speech.

Based on the above approaches, a number of characteristics of the speech output can be determined. This can include the volume, the pitch, the gender of the voice, the tone, the wait period between one word from the next, and other special emphasis on a word. This special emphasis can be some type of sound that is based on the semantic, but not the syntactic meaning of the word. Examples of the sound made can be a deep sigh, a grunt or a gasp. These sound-based expressions can convey a lot of meaning. Just as a picture is worth a thousand words, appropriate sound or emphasis provides additional meaning that can be very fruitful in any communication process.

The formatted text can be further modified to fit the requirements of the text-to-speech software engine. In one embodiment, the modification is through tagging, where a tag can be a command interpreted by the engine, and is not a word pronounced by the engine. The modified text is then fed to the speech hardware, which generates the speech output.

Note that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the inventive subject matter. Also, the features and advantages described in the specification are not all-inclusive. Other aspects and advantages of the present invention will become apparent to one of ordinary skill in the art, in view of the specification, which illustrates by way of example the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows one embodiment to implement the present invention.

FIG. 2 shows three approaches to transform a piece of input text into formatted text in the present invention.

FIG. 3 shows three approaches to transform a piece of input text based on the semantics of at least one word in the present invention.

FIG. 4 shows a number of characteristics of the speech output that can be determined in the present invention.

Same numerals in Figures 1-4 are assigned to similar elements in all of the figures. Embodiments of the invention are discussed below with reference to Figures 1-4. However, those skilled in the art will readily appreciate that the detailed description given herein with respect to these figures is for explanatory purposes as the invention extends beyond these limited embodiments.

DETAILED DESCRIPTION OF THE INVENTION

FIG. 1 shows one embodiment 100 to implement the present invention in a computing system. First a retriever 102 retrieves a piece of input text 104 previously entered into the system. The input text 104 can be in a storage medium. Then, a transformer 106 analyzes the input text 104 and transforms it into a formatted text 108. A modifier 110 further modifies this formatted text 108 to fit the requirements of a text-to-speech software engine 114 and a hardware engine 116 to generate the speech output 118.

The input text 104 belongs to a class that has at least one specific characteristic. For example, for some articles, the most important paragraphs are the first one and the last one, as in a newspaper. Another example is a weather report, as shown in the following example 1.

The formatted text 108 also has a number of characteristics. It can be independent of the text-to-speech software engine 114. In other words, the formatted text 108 can be written in a language that can be executed transparently in a number of different platforms. Then, the formatted text 108 can be further modified by the modifier 110 to fit the text-to-speech software engine 114.

In one embodiment, the formatted text 108 is written in Extensible Markup Language (XML), which is a data format for structured document interchange on the World Wide Web. XML is a standard method of document markup. One can define the grammar to mark-up this document in terms of tags and their attributes. A general description on XML can be found through the Web, with a URL of http://www.w3.org/XML, in an article entitled "Extensible Markup Language (XML)."

In another embodiment, the formatted text 108 includes tags, which define specific actions and can be implemented by subsequent engines that interpret those tags. Based on the XML example, an XML enabled browser can interpret the XML tags and carry out the appropriate actions as specified by the tags. The actions can include different audio rendering effects, such as background music, special effect sounds, and context sensitive sounds. For example, if the input text 104 is on waltz from Vienna, then Johann Strauss' Vienna Waltz might be broadcasted as background music while the text is read.

Other markup languages are also applicable to the present invention, such as:

(I) Standard Generalized Markup Language (SGML), as disclosed, for example, in *The SGML Handbook*, written by Charles Goldfarb, published in Clarendon Press, Oxford, in 1990.

(II) Spoken Text Markup Language (STML), as disclosed, for example, in *SSML: A Speech Synthesis Markup Language*, written by Paul Taylor and Amy Isard, published in Speech Communication 21, in 1996.

(III) *A Markup Language for Text-to-Speech Synthesis*, written by Richard Sproat, Paul Taylor, Michael Tanenblatt, and Amy Isard, published in the Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, in 1997.

FIG. 2 shows three approaches to transform a piece of input text 104 into formatted text 108. In one approach, the semantics 152 of the input text 104 is taken into account. One of the reasons why synthesized speech typically lacks the richness of a human voice is that the synthesizer does not understand the context of what it is reading. It is not source sensitive. In other words, it is not sensitive to the source. Imagine if you are reading about someone crying, probably your voice

would be different from your reading of someone laughing. A synthesizer reading both passages the same way would convey a wrong message. In another approach, the person hearing the speech output 118 is taken into account. One way to achieve this is by knowing the user profile 154 of that person. In yet a third approach, the hardware the user employs to listen to the speech output 118 is taken into account.

FIG. 3 shows three approaches to transform a piece of input text 104 based on the semantics 152 of at least one word in the text. In one approach, a person 175 determines the semantics 152. For example, the person, at strategic locations in the text, enters his inputs indicating how he prefers different words to be read. If he wants to read louder the sentence, "She loves me!", he would put an appropriate command character in the formatted text 108 at the end of the sentence. He transforms the input text 104 according to his preference.

In another approach, the transformation is done automatically by the transformer 106. It is through mapping some of the key words to a database 177. In one embodiment, the text is first parsed to identify specific words in the text, such as the proper nouns. These words are mapped to a pre-built database 177 to extract additional information. For example, if the proper noun is Microsoft, the word may be mapped to a database 177 describing different information on Microsoft, such as its current stock price.

In yet another embodiment, understanding the semantics is performed through an inference machine 179. Based on information in the text, the inference machine 179 determines the appropriate actions. For example, the inference machine 179 can perform natural language parsing, and statistical or rule-based parsing/understanding of the text. The rules can be created by experts in the field. The statistical approach can acquire information from training samples. Through

such parsing/understanding techniques, words are interpreted to determine subsequent actions. It should be obvious to those skilled in the art ways to perform natural language parsing, and statistical or rule-based parsing/understanding. Such techniques will not be further described in the present invention.

FIG. 4 shows a number of characteristics of the speech output 118 that can be determined in the present invention. Such determination is entered into the input text 104 to generate the formatted text 108. In one embodiment, the characteristics can be classified as prosodic characteristics, and include the volume 202, the pitch 204, and the tone 208. Other characteristics of the voice of the speech output 118 include the gender of the voice 206, and the perceived age 207 of the voice. The wait period 210 between one word from the next can also be modified. For example, after the phrase, "complete silence," nothing will be generated by the synthesizer for one second.

Another characteristic includes special emphasis 212 placed on a word. A special emphasis can be some type of sound that is based on the semantics of the word. Examples of such type of sound include a deep sigh, a grunt or a gasp. These sound-based expressions 214 can convey a lot of information. For example, a sigh after the sentence, "he smokes a cigarette," can illustrate relaxation.

Another output that can be determined is the ground frequency 216 of the speech output 118. It is the fundamental frequency contour of a piece of text when it is read out, and is typically not a constant value over an entire piece of text. However, as shown, for example, by the spectrogram of a segment of speech, one can estimate its principal frequency component through statistical

analysis. Such analysis should be obvious to those skilled in the art, and will not be further described.

The formatted text 108 is further modified to fit the requirements of a text-to-speech software engine 114. In one embodiment, the modification is through tagging. Here a tag can be a command tailored for the engine, and is not a word pronounced by the engine. The modified text 112 is then fed to the speech software and hardware to generate the speech output 118.

In another embodiment, the transformation to generate the formatted text 108 is based on one characteristic of the user listening to the speech output 118. This can be accomplished based on the user profile 154. For example, if the user is the chief executive of a company, probably she does not have that much time. Then only essential information will be communicated. If she is interested to find out today's temperature of San Jose, then instead of stating to her both the high, the low and the average temperature, the synthesizer only presents to her the average temperature--"The temperature of San Jose is 63 degrees." In another example, the user is hearing impaired, and the volume of the speech output 118 should be louder.

In yet another embodiment, the transformation to generate the formatted text 108 depends on the hardware engine 116, or the speech hardware, the user employs to listen to the speech output 118. For example, if the output is received by someone through a phone in a noisy room, probably, the volume of the speech output 118 should be louder. In this example, the phone with the room can be considered as the hardware engine 116.

In the above embodiments that are based on semantics 152, user profile 154 and hardware engine 116, the embodiments can be mixed and matched. For

example, the transformation can be based on the semantics 152 of at least one word of the text and the profile of the user listening to the speech output 118.

The above embodiments describe the transformer 106 and the modifier 110. In one embodiment, the input text 104 is changed only by the transformer 106, with the formatted text 108 tailored to one text-to-speech software and hardware engine 116. In yet another embodiment, the formatted text 108 is tailored to one specific text-to-speech hardware engine 116 without the need of a text-to-speech software engine. In a further embodiment, there is a text-to-speech engine coupled to the transformer 106 to generate the speech output 118.

Examples

The following two examples are taken from web pages from March 11, 1998. The original and the XML-formatted text are shown. The relevant information is extracted from the web pages using custom-built parsers and transducers. The final tagged text-to-speech input is generated using the syntax proposed by the Microsoft Speech SDK.

The first example is a weather forecast. The second example is a news story.

1. Weather Forecast

Source: (textual information extracted from html page http://weather.yahoo.com/forecast/-San_Jose_CA_US_f.html)

Yahoo! Weather - San Jose F° or C°

Weather : United States : California : San Jose

Today Thu Fri Sat Sun

63° [Image] Hi 74 [Image]	[Image]	[Image]	[Image]
at Mostly Partly	Partly		Partly
12:45pm Cloudy Lo 45 Cloudy	Cloudy	Showers	Cloudy
EST			
	Hi 68	Hi 60	Hi 64
Hi 66			
<101020 30405060 708090100+	Lo 48	Lo 56	Lo 50
Lo 53			

XML formatted:

```
<!DOCTYPE forecast http://weather.yahoo.com/...
<weather>
<region United States />
<state California />
<city San Jose />
<tempscale F />
<today temp="63 at 12:45pm EST" type="Mostly Cloudy", Hi=74,
Lo=45/>
<extendedForecast>
        <Thu type="Partly Cloudy", Hi=74, Lo=45 />
        <Fri type="Showers", Hi=60, Lo=56 />
        <Sat type="Partly Cloudy", Hi=64, Lo=50 />
        <Sun type="Partly Cloudy", Hi=66, Lo=53 />
        </extendedForecast>
</weather>
</forecast>
```

Tagged TTS input:

Weather forecast for San Jose , \pau=100\ California . \pau=500\ Today it is Mostly Cloudy \wav=mcloudy.wav\. The temperatures range between 74 and 45 degrees \Prn=Farenheit=farenhight\ . \pau=500\ Tomorrow it will be Partly Cloudy \wav=pcloudy.wav\, with temperatures between 68 and 48 degrees \Prn=Farenheit=farenhight\ .

We added the custom tag \wav\ to the set of tags of the Microsoft Speech SDK. \wav\ indicates that a sound file should be mixed in wherever this tag appears. The transformation rules from XML to the tagged data explicitly leave out information here. The current temperature and the extended forecast information are not spoken in this example.

2. News

Source: (extract from http://dailynews.yahoo.com/headlines/top_stories/story.html?s=n/-reuters/980311/news/stories/weather_13.html)

```
<table> <tr> <td>
<title>Many Without Power in Midwest; South Colder</title>

<hr>
<strong>
<!-- Yahoo TimeStamp: 889598580 -->
Wednesday March 11  1:43 AM EST
</strong>
<h2>Many Without Power in Midwest; South Colder</h2>

<!-- Yahoo TimeStamp: 889598580 -->

<p>
        By Andrew Stern
<p>
        CHICAGO (Reuters) - Temperatures plunged Tuesday in the wake of a late winter storm that knocked out power to hundreds of thousands of people across the Midwest and triggered flooding in the Southeast.
        <p>
        &amp;quot;Several counties have declared snow emergencies, meaning people should stay off the roads,&amp;quot; Indiana Emergency Management spokesman Alden Taylor said. &amp;quot;There are so many cars stranded on the roads, it's hard for plows to get through.&amp;quot;
<p>
                ...
```

XML formatted:

<!DOCTYPE news http://dailynews.yahoo.com/headlines ...
<http://dailynews.yahoo.com/headlines?.. > >
<headl> Many Without Power in Midwest; South Colder </headl>

<author> By Andrew Stern </author>

<place> Chicago </place> <agency> (Reuters) </agency> -
<main> Temperatures <stress> plunged </stress> Tuesday in the wake of a late winter storm that knocked out power to <stress> hundreds of thousands of people </stress> across the Midwest <subclause> and triggered flooding in the Southeast </subclause>.

<quote male Alden-Taylor> Several counties have declared snow emergencies, <subclause> meaning people should stay off the roads </subclause>, </quote><subclause> Indiana Emergency Management spokesman Alden Taylor said </subclause>.
<quote male Alden-Talyor> There are so many cars stranded on the roads, <2ndclause> it's hard for plows to get through </2ndclause>.</quote>
</main>
</news>

Used formatting rules: (a) mark all standard elements of an article such as headline, author, etc. (b) identify the syntactic structure of the text (e.g. subclauses), (c) find quotes, (d) find verbs which stress events, and (e) mark phrases that emphasize exceptional properties of events.

Tagged TTS input:

\wav=ping.wav\\Vce=Language=English,Gender=male,Sytle=Business\
Many Without Power in Midwest; South Colder
\pau=1000\

\wav=dong.wav\Chicago \pau=500\ -
\Vce=Language=English,Gender=female,Sytle=Business\
Temperatures \emp\plunged Tuesday in the wake of a late winter storm ,
\pau=100\ that knocked out power to \emp\ hundreds-of-thousands-of-people across the Midwest , and triggered flooding in the Southeast.

\Vce=Language=English,Accent=Midwest,Gender=male,Age=40\
\quot\ Several counties have declared snow emergencies, \pau=100\ meaning people should stay off the roads, \quot\ \Rst\ Indiana Emergency Management spokesman Alden Taylor said.
\Vce=Language=English,Accent=Midwest,Gender=male,Age=40\
\quot\ There are so many cars stranded on the roads, it's hard for plows to get through.\quot\ \Rst\

Other embodiments of the invention will be apparent to those skilled in the art from a consideration of this specification or practice of the invention disclosed herein. It is intended that the specification and examples be considered as exemplary only, with the true scope and spirit of the invention being indicated by the following claims.

CLAIMS

1. A method of synthesizing speech from a piece of input text comprising the steps of:

retrieving the input text entered into a computing system; and

transforming the input text based on the semantics of at least one word of the input text to generate a formatted text for speech synthesis.

2. A method as recited in Claim 1 wherein the step of transforming is performed by an author who transforms the input text to the formatted text.

3. A method as recited in Claim 1 wherein the step of transforming is performed through mapping the at least one word to one or more entries in a database.

4. A method as recited in Claim 1 wherein the step of transforming is performed through an inference machine, which infers an action based on the at least one word.

5. A method as recited in Claim 1 wherein:

the formatted text can be coupled to more than one type of text-to-speech software engines to produce the speech output; and

the formatted text is independent of the text-to-speech software engines.

6. A method as recited in Claim 1 wherein the input text belongs to a class that has at least one specific pattern.

7. A method as recited in Claim 1 wherein the semantics define an audio rendering effect.

8. A method as recited in Claim 1 wherein the formatted text is written in XML.

9. A method as recited in Claim 1 wherein when synthesized, the volume of the at least one word is determined in view of the semantics.

10. A method as recited in Claim 1 wherein when synthesized, the pitch of the at least one word is determined in view of the semantics.

11. A method as recited in Claim 1 wherein when synthesized, the gender of the voice to pronounce the at least one word is determined in view of the semantics.

12. A method as recited in Claim 1 wherein when synthesized, the perceived age of the voice to pronounce the at least one word is determined in view of the semantics.

13. A method as recited in Claim 1 wherein when synthesized, a prosodic characteristic of at least one word is determined in view of the semantics.

14. A method as recited in Claim 1 wherein when synthesized, the tone of the at least one word is determined in view of the semantics.

15. A method as recited in Claim 1 wherein:

there is a period of silence between the reciting of the at least one word, and the word after the at least one word; and

the period is determined in view of the semantics.

16. A method as recited in Claim 1 wherein when synthesized, the at least one word is pronounced with special emphasis, which is determined in view of the semantics.

17. A method as recited in Claim 1 wherein when synthesized, the at least one word is pronounced with an additional sound-based expression, that is not based on the syntactic, but on the semantics of the at least one word.

18. A method as recited in Claim 1:

the formatted text is coupled to a text-to-speech software engine; and

the method further comprises the step of modifying the formatted text following one or more characteristics of the text-to-speech software engine.

19. A method as recited in Claim 18 wherein the step of modifying includes the step of tagging the formatted text, with a tag being a command tailored for the text-to-speech software engine.

20. A method as recited in Claim 1 wherein the step of transforming also depends on at least one characteristic of the user listening to the synthesized speech.

21. A method as recited in Claim 20 wherein the at least one characteristic is that the user is hearing impaired.

22. A method as recited in Claim 20 wherein the at least one characteristic is that the user is visually impaired.

23. A method as recited in Claim 1 wherein the step of transforming also depends on at least one characteristic of the speech hardware a user employs to listen to the synthesized speech.

24. A computing apparatus for synthesizing speech from a piece of input text comprising:

a retriever configured to retrieve the input text entered into the computing apparatus; and

a transformer configured to transform the input text based on the semantics of at least one word of the input text to generate a formatted text for a text-to-speech engine.

25. A method of synthesizing speech from a piece of input text comprising the steps of:

retrieving the input text entered into a computing system; and

transforming the input text based on at least one characteristic of the user listening to the synthesized speech to generate a formatted text for a text-to-speech software engine.

26. A method as recited in Claim 25 wherein the step of transforming also depends on the semantics of at least one word of the input text.

27. A method of synthesizing speech from a piece of input text comprising the steps of:

retrieving the input text entered into a computing system; and

transforming the input text based on at least one characteristic of the speech hardware a user employs to listen to the synthesized speech, to generate a formatted text for a text-to-speech software engine.

28. A method as recited in Claim 27 wherein the step of transforming also depends on the semantics of at least one word of the input text.

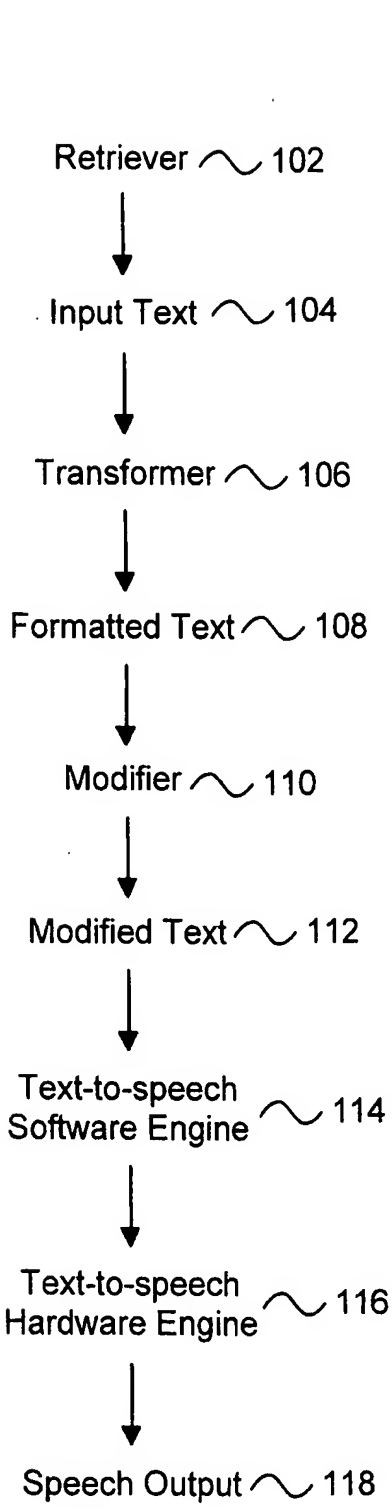

100
Retriever 102
Input Text 104
Transformer 106
Formatted Text 108
Modifier 110
Modified Text 112
Text-to-speech Software Engine 114
Text-to-speech Hardware Engine 116
Speech Output 118

Figure 1

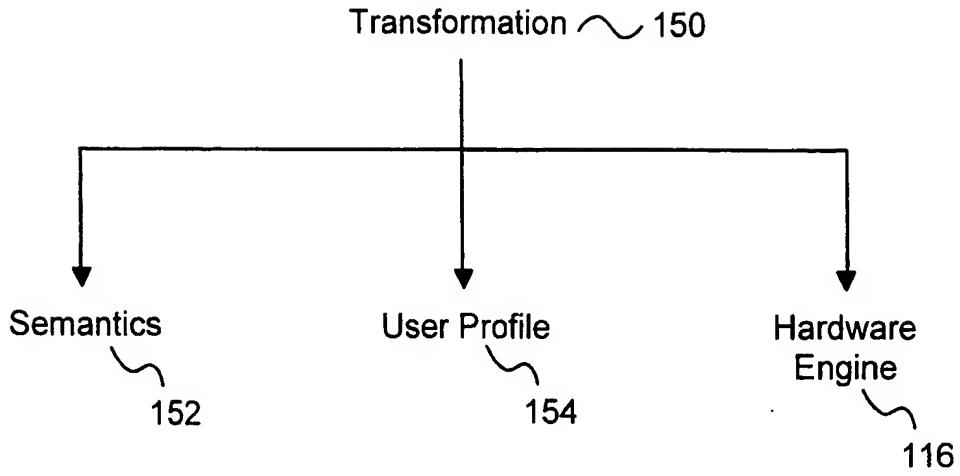

Transformation
150
Semantics
152
User Profile
154
Hardware
Engine
116

Figure 2

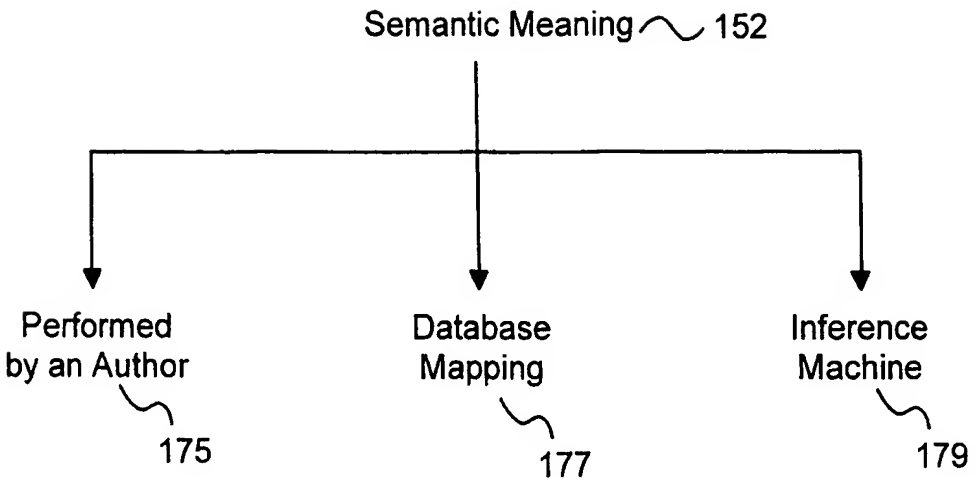

Semantic Meaning
152
Performed by an Author
175
Database Mapping
177
Inference Machine
179

Figure 3

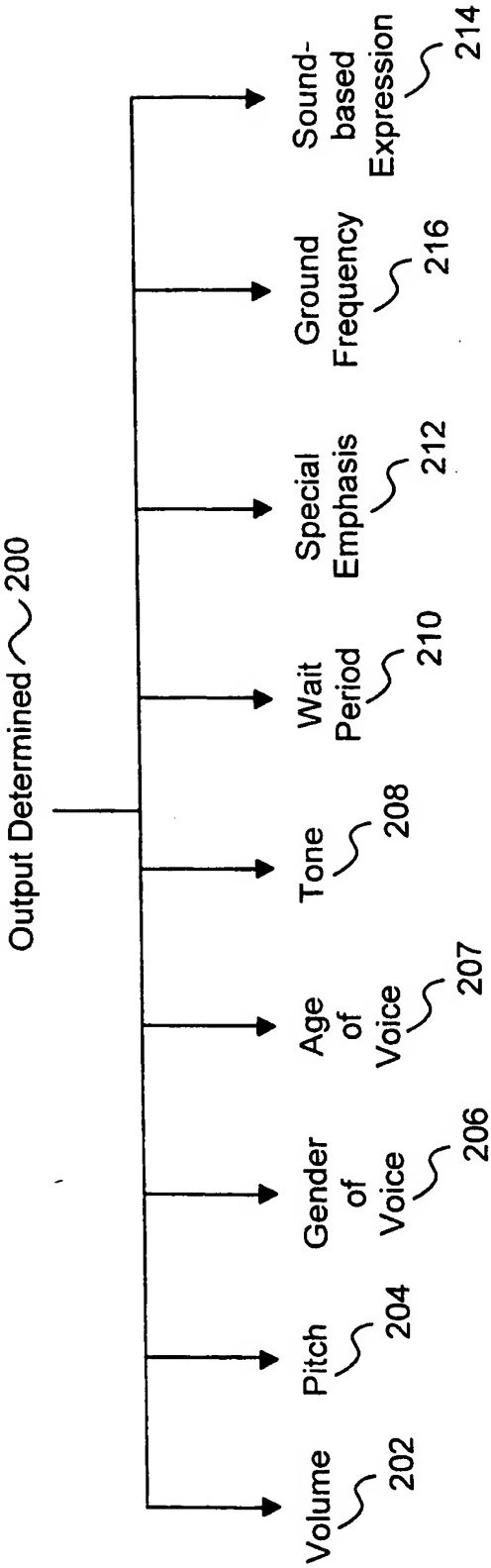

Output Determined
200
Volume
202
Pitch
204
Gender of Voice
206
Age of Voice
207
Tone
208
Wait Period
210
Special Emphasis
212
Ground Frequency
216
Sound-based Expression
214

Figure 4

INTERNATIONAL SEARCH REPORT

International Application No
US 99/13329

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G10L5/04

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 761 640 A (KALYANSWAMY ASHOK ET AL) 2 June 1998 (1998-06-02) abstract; claim 4	1,6,18, 24
X	WO 96 22594 A (CENTIGRAM COMMUNICATIONS CORP) 25 July 1996 (1996-07-25) abstract; claim 1; figure 3	1,6,18, 24
Y		10,12,20
Y	US 5 029 214 A (HOLLANDER JAMES F) 2 July 1991 (1991-07-02) abstract; claim 1	20
Y	EP 0 841 625 A (SOFTMARK LIMITED) 13 May 1998 (1998-05-13) abstract; claims 1,10	1,6,10, 12,18, 20,24

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
7 October 1999	15/10/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Van Doremalen, J

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
T/US 99/13329

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	DATABASE WPI Week 199620 Derwent Publications Ltd., London, GB; AN 1996-192669 XP002118067 & JP 08 063188 A, 8 March 1996 (1996-03-08) abstract -----	1,6,10. 12,18. 20,24

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
US 99/13329

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5761640 A	02-06-1998	NONE	
WO 9622594 A	25-07-1996	US 5634084 A	27-05-1997
		AU 4602696 A	07-08-1996
US 5029214 A	02-07-1991	NONE	
EP 0841625 A	13-05-1998	EP 0841624 A	13-05-1998
		AU 4438197 A	14-05-1998
		CA 2220314 A	08-05-1998
		IE 79053 B	08-04-1998
		JP 10143485 A	29-05-1998
JP 8063188 A	08-03-1996	JP 2770747 B	02-07-1998
		US 5857170 A	05-01-1999